Communication-Efficient Edge AI: Algorithms and Systems

Year: 2020 | Citations: 383 | Authors: Yuanming Shi, Kai Yang, Tao Jiang, Jun Zhang, K. Letaief

---

Abstract

Artificial intelligence (AI) has achieved remarkable breakthroughs in a wide range of fields, ranging from speech processing, image classification to drug discovery. This is driven by the explosive growth of data, advances in machine learning (especially deep learning), and the easy access to powerful computing resources. Particularly, the wide scale deployment of edge devices (e.g., IoT devices) generates an unprecedented scale of data, which provides the opportunity to derive accurate models and develop various intelligent applications at the network edge. However, such enormous data cannot all be sent to the cloud for processing, due to the varying channel quality, traffic congestion and/or privacy concerns, and the enormous energy consumption. By pushing inference and training processes of AI models to edge nodes, edge AI has emerged as a promising alternative. AI at the edge requires close cooperation among edge devices, such as smart phones and smart vehicles, and edge servers at the wireless access points and base stations, which however result in heavy communication overheads. In this paper, we present a comprehensive survey of the recent developments in various techniques for overcoming these communication challenges. Specifically, we first identify key communication challenges in edge AI systems. We then introduce communication-efficient techniques, from both algorithmic and system perspectives for training and inference tasks at the network edge. Potential future research directions are also highlighted.